Language and Biology

In-Silico Genomic Biology

What is the “genomic excitement” about?

http://www.tigr.org/tdb/tdb.html

The entire DNA sequences of important biological organisms are now available to the research community.

What is next?

Manuals of Life: Producing integrated databases of biological knowledge.

Bio-interfaces: Developing languages to express knowledge about biological data, processes and experiments.

Virtual Cells: Producing models of biological cells that are consistent with genomic data.

Bio-programming: Creating systems for manufacturing predictable biological systems that are controlled and monitored by a semi-automated in-silico bio-interfaces and virtual cells.

* these research tasks are listed in order of estimated difficulty.

What can we do now:

Engineering Inspired Representations and Algorithms :

- Gene Finding with HMMs
- Protein Function Assignment with Probabilistic Representations (PFAM)
- tRNA scan
- Boolean Network Modeling of Gene Regulation
- Gene Fusion Events by Database Search
- Discovery of Co-regulation by Text Search
- Bio-Spice
- MicroArrays
- Experimental Design (e.g, multiplex PCR)
- Fusion of Information Sources (Eisenberg)

Gene Finding Observation

- Most of the predicted genes in currently available genomes were predicted by hidden Markov models interpolated Markov models, and edit-distance technologies originally developed by the speech recognition community.
- Gene recognition technology in microbial DNA is achieving 98% accuracy, a remarkably low error rate for predictive biology systems.
- **(See our system Glimmer, http://www.tigr.org).**

These techniques were developed by the DARPA speech/language understanding initiative.

TALKS IN OUR SESSION

- Whole Genome Analysis (Salzberg)
  - Detecting Foreign DNA – important for detecting possibly new microbial organisms.
  - Tuberculosis → Leprosy inversion + deletions – important for building and detecting new microbial organisms.
- Predictive Biology (Stormo) — a step towards automating the construction of DNA binding proteins.
- Bio-Spice – Multi-level reasoning about Biological Systems (Arkin) – a step towards virtual cells.
- A language to describe gene interaction (Brent).

More Talks and Ideas

- Language for describing gene structure and protein structure.
- Sophisticated analysis of medical literature correlated in a non-trivial manner to results of experiments .
- Language for describing protein structure and activity for function prediction.
- Vertical integration of information sources and data types for constructing on-line bio-manuals.
- Virtual cell

A pragmatic methodology:

Computationally well-formed biological questions (protein shape, multiple sequence alignment, motif detection, evolutionary trees, …

How do algorithms scale up or perform in the presence of increasing data – benchmarks, performance analysis…

A high-risk methodology:

Not well formed problems: (virtual cell, pathways, gene function, semi-automated genome comparisons, building genetic switches,…)

How to bring different perspectives to make progress in moving towards effective algorithms – very important to make big advances.

What should DARPA do?

Computational Technique Oriented or Biological Problem Oriented?

•Focus on few computational techniques and representations such as Probabilistic Networks, Logical Rules, Dynamical Systems, Grammatical Representations, Language Processing, Approximation Algorithms for Hard Optimization Problems

•A cut across the most creative ideas to model and possibly control a complex biological system such as the immune system, human pathogens - virtual cells, important biological pathways that cut across species, genetic switches.

Requires a cut across theory and experiments!